VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Multi-state models for clustered duration data: an application to workplace effects on individual sickness absenteeism

Lindeboom, M.; Kerkhofs, M.

1998

document version
Early version, also known as pre-print

Link to publication in VU Research Portal

citation for published version (APA)
Lindeboom, M., & Kerkhofs, M. (1998). *Multi-state models for clustered duration data: an application to workplace effects on individual sickness absenteeism*. (Research Memorandum; No. 1998-8). FEWEB.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 04. May. 2023

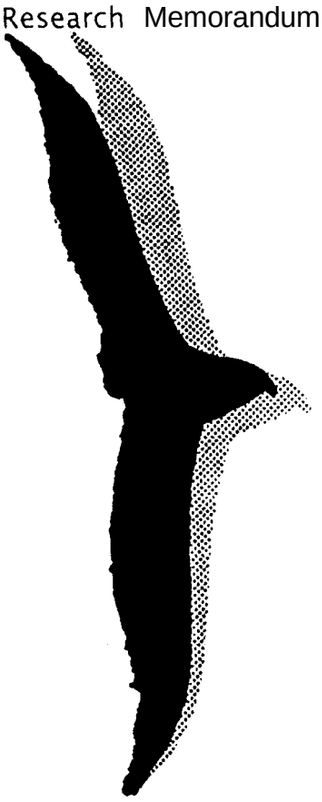
Faculteit der Economische Wetenschappen en Econometrie

Serie research memoranda

Multi-s tate Models for Clustered Duration Data: an application to workplace effects on individua.l sickness absenteeism

Maarten Lindeboom
Marcel Kerkhofs

Research Memorandum 1998-8 March 1998



/a/l/e/r/t/

applied
labour
economics
research
team

vrije Universiteit *amsterdam*

MULTI-STATE MODELS FOR CLUSTERED DURATION DATA

AN APPLICATION TO WORKPLACE EFFECTS ON INDIVIDUAL SICKNESS ABSENTEEISM

Maarten Lindeboom

Marcel Kerkhofs

BIBL. VRIJE UNIVERSITEIT AMSTERDAM
ET. 08348

March 1998

Marcel Kerkhofs is at the Department of Economics of Leiden University, The Netherlands. Maarten Lindeboom is at the-Department of Economics of the Free University, the Tinbergen Institute and the Economics Institute of Tilburg University, The Netherlands. Part of this work was done while both authors were at Leiden University. The authors acknowledge the valuable comments and suggestions of Leo Aarts, Gerard van den Berg, John Ham, Jim Heckman, Peter Kooreman and Michael Visser. Robert Moffit and two anonymous referees of this journal in particular gave many helfull suggestions that have led to considerable improvement of this paper. Correspondence address: Department of Economics, Free Universitya, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.

MULTI-STATE MODELS FOR CLUSTERED DURATION DATA - AN APPLICATION TO WORKPLACE EFFECTS ON INDIVIDUAL SICKNESS ABSENTEEISM

Maarten Lindeboom

Marcel Kerkhofs

Abstract

In this paper we specify and estimate three state duration models of work, sickness and exit from the job to explain individual absenteeism behaviour of primary school teachers. There is a large variation of sickness absenteeism records across schools and absenteeism records of workers within a school appear to be related. This clustering of individual absenteeism data may to a large extend be caused by workplace effects. Since it will be difficult to fully capture workplace effects with observed characteristics of the workplace, we also account for unobserved workplace effects in the models. The most flexible specification allows for non-parametric baseline hazards that differ per exit rate and workplace. A stratified partial likelihood approach is used to estimate the regression coefficients of this model. Conditional on these estimates we recover fixed unobserved workplace effects and semi-parametric baseline hazards in order to detect the causes for the observed variation and clustering in the data.

Keywords: Multi-state duration models, clustered duration data, fixed effects, Concentrated likelihood, stratified Partial Likelihood, Sickness Absenteeism.

.

1. INTRODUCTION

We specify and estimate fixed effect multi-state duration models to explain individual absenteeism behaviour of primary school teachers in The Netherlands. Sickness absence records of employees in this sector exceed the averages of most other sector and absenteeism across schools ranges from schools with a few spells lasting single days ('healthy schools) to schools with a high number of spells lasting several weeks if not months ('sick' schools). Moreover, the larger part of working days lost due to sickness absenteeism is concentrated at a relatively small number of schools. This clustering of absenteeism records may be determined by specific individual circumstances and/or circumstances specific to the work environment. The circumstances specific to the work environment depend on the job requirements, specifics of the sickpay scheme, age and quality of the buildings, the quality of the pupils (ethnic origin, native language, social background), norms or moral attitude towards absence behaviour, quality of the management, policies towards prevention of absence behaviour etc. Specific individual circumstances may account for a clustering of absenteeism records within a school if sickness prone people are assigned to specific schools. We denote this as a sorting effect. It may be clear that for policy purposes it is important to distinguish between the different causes of absenteeism, and their importance in explaining observed patterns.

We focus on sickness incidence and sickness duration of individual teachers within a school to assess whether sorting effects or workplace characteristics cause the large variance and the clustering in the data. A natural way to do this is in the context of a multi-state duration model. Special attention is paid to the role workplace effects in a multi-state model as these may be of prime importance for the observed clustering. In general it will be very difficult to fully capture the influence of work environment with observed workplace characteristics and it is well known that parameter estimates of duration models are biased if heterogeneity is not adequately accounted for (see e.g. Lancaster (1990)). We therefore specify models that take account of unobserved workplace effects in a flexible way. As in the

epidemiological literature on related failure times within a household (see e.g. Clayton (1978), Ridder & Tunali (1989) and Gönöl & Srinivasan (1993)), sickness and work spells of individuals within a school may be related by common unknown factors. The most flexible specification allows for non-parametric baseline hazards that differ per exit rate and school. A stratified partial likelihood approach is used to estimate the regression coefficients of this model. We show that this stratified partial likelihood can be derived using a concentrated likelihood approach. This concentrated likelihood approach allows us to recover estimates of unobserved workplace effects and non-parametric baseline hazards given estimates of the regression coefficients obtained from the stratified partial likelihood. The unobserved workplace effects are used to detect the causes for the observed variation and clustering in the absenteeism records.

In the analyses we find strong effects of both observed personal characteristics and school • characteristics. From a comparison of a range of models we conclude that it is important to allow for unobserved workplace/school effects, but that this also needs to be done in the most flexible way. Unobserved workplace specific effects account to a large extent for the observed variation of sickness absenteeism across schools. We also find that the observed clustering in 'healthy' schools and 'sick' schools is a result of unobserved school effects instead of a teacher sorting effect. In an additional analysis we relate the school specific fixed effects to a range of observed exogenous school variables. The estimates indicate that the school specific effects are hardly related to the exogenous variables of the type available in the data. It remains however, that workplace effects are important in explaining sickness absence patterns, and a better understanding of these workplace conditions will prove to be essential in reducing sickness absenteeism.

The remainder of the paper is organized in 4 sections. Statistical models for clustered duration data are presented in section 2. Section 3 gives a brief description of our sample of primary school teachers that we use in the application. Institutional features of the educational system in the Netherlands are important to understand sickness absenteeism. We give a brief desription in section 3. Section 4 contains three subsections. Subsection 4.1 presents the

empirical implementation. Estimation results are discussed in subsection 4.2. This subsection also contains a comparison of the performance of a range of alternative models that we have estimated. In subsection 4.3 we pay special attention to the effect of (unobserved) school specific effects and the role they play in explaining observed absenteeism patterns. Section 5 concludes.

3 STATISTICAL MODELS FOR CLUSTERED DURATION DATA

We focus on two dimensions of sickness absenteeism: sickness incidence and sickness duration. A natural way to model this is in the context of a multi-state duration model. An individual worker (indexed by i) at a workplace/cluster (indexed by m) can either be at work (W) or sick (S). Individual workers are allowed to leave the job (E). Let's for now assume that we observe complete histories of work and sickness absence of individual workers, i.e. we observe individuals from the moment that they enter the job up to the moment that they leave the job. We discuss sampling issues in section 4.1. The exit state is denoted by (E). Consequently, a spell of sickness may either end in a work spell or in an exit out of the job. In accordance with this we define $\theta^{S,W}$ as the exit rate for a transition from sickness to work, and $\theta^{S,E}$ as the transition rate from sickness to out of a job. Similarly, a work spell may end in a sickness spell or in an exit out of the job and $\theta^{W,S}$ and $\theta^{W,E}$ are the exit rates associated with these transitions. We take the transition rates to be of the mixed proportional hazard (MPH) type, and (suppressing the index for individual variation, i) write them as:

$$\theta^{K,L}(t;x,\eta_m^{K,L},\beta^{K,L}) = \theta_0^{K,L}(t,\eta_m^{K,L})\,\theta_1^{K,L}(x;\beta^{K,L}), \qquad K \neq L \tag{1}$$

We refer to Lancaster (1990) for a theoretical exposition on MPH models. K and L ($K,L \in \{S,W,E\}$), refer to the state of origin and destination, respectively and t is the waiting time. The term η_m^{KL} is unobserved and specific to a cluster m, $m = 1,\ldots,M$, and may differ for

each of the hazard rates that we consider. In principle the baseline hazard θ_0 is an arbitrary function of unobserved cluster-specific heterogeneity and duration dependence. The regression function θ_1 includes a vector of observed **characteritics** x. The vector x may include observed individual characteristics as well as observed characteristics of the cluster. For ease of exposition we take x time constant, though this assumption can be relaxed without altering the results presented below.

If we assume that all individual differences can be described by x and $\eta_m^{K,L}$, $K,L \in \{ S,W,E \}$, $K \neq L$, $m=1,..,M$, then, conditional on these factors, the individual failure times in each of the states can be treated as independent and the total likelihood function factorizes in seperate parts, each associated with one of the exit rates that we consider. For instance the likelihood for the transition from sickness to work may be written as:

$$\mathcal{L}^{S,W} = \prod_{i=1}^{N} \theta^{S,W}(t_i;x_i,\eta_{m(i)}^{S,W})^{\delta_i^{S,W}} \exp\{ - \int_0^{t_i} \theta^{S,W}(u;x_i,\eta_{m(i)}^{S,W})\, du \} \qquad (2)$$

The scalar $\delta^{S,W}_i$ is an indicator that equals 1 if a spell in state S ends with a transition to state W and $\eta_{m(i)}^{K,L}$ is i's school specific fixed effect. Assumptions regarding the baseline hazard $\theta_0^{K,L}(t,\eta_m^{K,L})$ to a large extend determine the estimation strategy.

2.1 *A model with fixed unobserved workplace effects*

As in Gönül & Srinivasan (1993) one could specify the baseline hazard in (1) as the product of a function for the duration dependence $\theta_0^{K,L}(t)$ and a time constant unobserved term $\eta_m^{K,L}$:

$$\theta_0^{K,L}(t,\eta_m^{K,L}) = \theta_0^{K,L}(t)\, \eta_m^{K,L} \qquad K,L \in \{ S,W,E \},\ K \neq L,\ m=1,..,M \qquad (3)$$

In a random effect specification of the unobserved components one assumes these terms to be generated by some specified multi-dimensional distribution, that have to be integrated out of the likelihood function. A disadvantage of this approach is that, due to the dependency of the unobserved components, the likelihood fails to factorize. This will make estimation cumbersome, especially when one whishes to estimate the baseline hazards non-parametrically

(as we will do in subsection 2.2 and 2.3). Moreover, a random effect approach requires the terms $\eta_m^{K,L}$ to be independent of (other) included regressors x. This assumption may easily be violated in practical situations where one samples individuals at a point in time.

Alternatively one could follow a fixed effects approach that treats $\eta_m^{K,L}$ as unknown parameters that need to be estimated along with the other parameters. An advantage of this approach is that $\eta_m^{K,L}$ need not be orthogonal to x, and that the likelihood remains very simple and still factorizes in different parts for each transition rate.

Likelihood terms like (2) contain a set of M nuisance parameters $\eta_m^{S,W}$ and consistency of the maximum likelihood estimates depends upon the implied role of asymptotics in the model. Consistency is for instance obtained if we rely on asymptotics in time or in the number of individuals. This guarantees that sample information grows over time for a fixed number of parameters and the usual properties of the maximum likelihood estimator apply.

Joint estimation of the M cluster/workplace specific effects in (2) along with the other parameters would lead to enormous computational problems. As in Gönül & Srinivasan (1993) we can therefore follow an approach that concentrates the workplace specific fixed effects out of the likelihood. Substitution of (3) into (2) and taking the first derivative with respect to $\eta_m^{S,W}$ of the logarithm of (2) and equalizing this expression to zero, one obtains:

$$\hat{\eta}_h^{S,W} = \frac{\sum_{\substack{i=1 \\ m(i)=h}}^{N} \delta_i^{S,W}}{\sum_{\substack{i=1 \\ m(i)=h}}^{N} \int_0^{t_i} \theta_0^{S,W}(u)\,du\ \theta_1^{S,W}(x_i;\beta^{S,W})} \tag{4}$$

Substitution of (4) in the logarithm of (2) gives the concentrated likelihood function:

$$\mathcal{L}_c^{S,W} \propto \prod_{i=1}^{N} \left(\frac{\theta_0^{S,W}(t_i)\ \theta_1^{S,W}(x_i;\beta^{S,W})}{\sum_{i=1}^{N} \int_0^{t_i} \theta_0^{S,W}(u)\,du\,\theta_1^{S,W}(x_i;\beta^{S,W})} \right)^{\delta_i^{S,W}} \tag{5}$$

Expressions for the other transition rates are analogous. Likelihood (5) is a simple likelihood

function that needs to be optimized with respect to $\beta^{S,W}$ and the parameters of the baseline hazard $(\theta_0^{S,W}(t))$. Under the usual regularity conditions, consistent estimates are provided. Next given these estimates, equation (3) could be used to obtain estimates of the workplace specific effects $(\hat{\eta}_m^{S,W})$. Note that $\hat{\eta}_m^{S,W}$ is set equal to zero for companies where no failures take place. In practice this means that in the estimation of (5), these observations do not contribute and that hazard rates of these clusters is set to zero.

The multi-state model without unobserved workplace specific effects is nested in the fixed effect specification and follows from the restriction that $\eta_m^{S,W}=\eta_{m'}^{S,W}$, $\forall\ m,m' = 1,\ldots,M$. Consequently, simple likelihood ratio tests could be employed to test for the relevance of unobserved cluster effects. We return to this issue by the end of this section.

Likelihood contributions like (5) are convenient, as they are simple, and unobserved workplace effects are allowed for in a straightforward way. A disadvantage is that the baseline hazards $(\theta_0^{K,L}(t)$ are estimated jointly with $\beta^{K,L}$ and therefore requires a priori, possibly restrictive parametric assumptions. A way relax the restrictiveness is to use partial likelihood methods that acknowledge unobserved workplace effects.

2.2 *Partial likelihood, Non-parametric baseline hazards and fixed unobserved workplace effects*

Contributions to the partial likelihood function are based on the conditional probability that a spell i ends, given the riskset $R_i^{K,L}$, defined as the set of spells having the same duration as i or longer. This conditional probability is a simple ratio of the hazard rate of i relative to the sum of all individuals that are exposed to the risk. As a consequence, due to the proportionality assumption of the hazard, factors common to all individuals cancel from the expression. So, with the baseline hazard specified as in (3) the partial likelihood associated with a transition from state K to state L becomes:

$$\mathcal{L}_{pl}^{K,L} = \prod_{i=1}^{N} \left(\frac{\theta_1^{K,L}(x_i;\beta^{K,L})\ \eta_{m(i)}^{K,L}}{\sum_{j \epsilon R_i^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L})\ \eta_{m(j)}^{K,L}} \right)^{\delta_i^{K,L}} \tag{6}$$

$\delta_i^{K,L}$ is an indicator that equals 1 if i is observed to make a transition from state K to L. *The* expression for the likelihood implies that for the estimation of the regression function and the workplace specific fixed effects, the baseline hazards can be left unspecified. The partial likelihood (6) may be flexible with respect to θ_0, but may still be cumbersome to optimize as still M fixed effects need to be optimized along with the other parameters. A way to circumvent this problem is to concentrate the logarithm of (6) with respect to the fixed effects to obtain:

$$\eta_m^{K,L} = \frac{\sum_{i \epsilon R_{m,j}^{K,L}} \delta_i^{K,L}}{\sum_{i=1}^{N} \delta_i^{K,L} \frac{\sum_{j \epsilon R_{m,j}^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L})}{\sum_{j \epsilon R_i^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L})\ \eta_{m(j)}^{K,L}}} \tag{7}$$

$R^{K,L}{}_{m,i}$ is the risk set of spells at school m having the same duration as i or longer. According to (7) the workplace specific fixed effect of school m is the sum of scores at that school divided by a weighted average of the scores of all schools. Unfortunately (7) does not provide a closed form solution for $\eta_m^{K,L}$, so that these can not be concentrated out of the partial likelihood (6). Therefore, a procedure must be applied in which in each iteration of the maximization procedure, equation (7) is used iteratively to solve the fixed effects. This procedure is computationally more demanding than direct optimization of the concentrated likelihood (5).

Given estimates of $\beta^{K,L}$ and $\eta_m^{K,L}$ the non-parametric baseline hazards could be recovered, using the (concentration) technique suggested by Breslow (1974). In this approach the non-parametric baseline hazard $\theta_0^{K,L}(t)$ is a piecewise constant function with discontinuities at each observed failure point. Likelihood (2) could be rearranged to:

$$\mathcal{L}^{K,L} = \prod_{i=1}^{N} \{ \theta_0^{K,L}(t_i) \}^{d_i^{K,L}} \{ \prod_{j \in D_i^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L}) \eta_{m(j)}^{K,L} \} \exp\{ - \theta_0^{K,L}(t_i) \sum_{j \in R_i^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L}) \eta_{m(j)}^{K,L} \} \qquad (2')$$

With $d_i^{S,W}$ as the number, and $D_i^{S,W}$ as the set of individuals that experience a transition from K to L at transition time t_i. Maximization of the logarithm of (2') with respect to $\theta_0^{K,L}(t)$ gives the step function:

$$\hat{\theta}_0^{K,L}(t) = \frac{d_i^{K,L}}{\sum_{j \in R_i^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L}) \eta_{m(j)}^{K,L}} \quad , \qquad i = \underset{\{j \mid t_j < t\}}{arg\max}\ t_j \qquad (8)$$

According to (8), $\hat{\theta}_0^{K,L}(t)$ could be viewed as a Kaplan-Meier estimate of the hazard after proper reweighting of the data. Substitution of (8) into (2') gives a likelihood function that is proportional to (6). So, estimates of the partial likelihood (6) and the concentrated likelihood produce identical results. Hence, estimates of $\beta^{K,L}$ and $\eta_m^{K,L}$ obtained from (6) and (7) could be used in (8) to calculate the non-parametric baseline hazard.

The partial likelihood (6) nests the partial likelihood without unobserved fixed effects by imposing the restriction $\eta_m^{S,W} = \eta_{m'}^{S,W}$, $\forall\ m,m' = 1,\ldots,M$. So in principle simple likelihood ratio tests could be applied to test for the relevance of unobserved workplace effects in a (relatively) flexible model. It should be noted, however, that schools with only censored durations effectively do not contribute to likelihood (6). Consequently, models with unobserved cluster effects are estimated on a different sample than more traditional models that do not allow for unobserved cluster effects. Therefore, as an alternative, also a Hausman tests could be performed to test for the relevance of school specific fixed effects.

The model presented above is appealing as duration dependence is allowed for in the most flexible way. However, it may still be restrictive, as the partial likelihood (6) depends on the assumption that the baseline hazard factorizes in two seperate parts and that workplace specific effects are constant over time. Moreover, the estimation procedure is computationally quite demanding.

2.3 *Stratified partial likelihood, Non-parametric workplace specific baseline hazards*

An alternative to the model in subsection 2.2 is a stratified partial likelihood model such as in Ridder & Tunali (1989, 1990). In this approach the baseline function is treated as an arbitray function of t and $\eta_m^{K,L}$, i.e. no specific functional form is imposed on the baseline hazard in (1). The stratified partial likelihood approach stratifies the risks sets into different subsets, each belonging to a seperate cluster. As may be intuitively clear, cluster effects are not relevant in a comparison of individuals belonging to the same cluster, and therefore cancel from the expression for the likelihood. The stratified partial likelihood of a transition from state K to L is given by:

$$\mathcal{L}_{spl}^{K,L} = \prod_{i=1}^{N} \frac{\theta_1^{K,L}(x_i;\beta^{K,L})^{\delta_i^{K,L}}}{\sum_{j \in R_{m,i}^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L})} \tag{9}$$

Similar to the previous subsection Breslow type of concentration arguments could be used to obtain the stratified partial likelihood (9) from a likelihood rearranged like (2')[1]. The procedure is particularly convenient, because it allows us to recover estimates of the non-parametric workplace/school specific baseline hazard, given estimates of the regression parameters from (9). In particular:

$$\hat{\theta}_{0,m}^{K,L}(t) = \frac{d_{i,m}^{K,L}}{\sum_{j \in R_{i,m}^{K,L}} \theta_1^{K,L}(x_j;\beta^{K,L})}, \qquad i = \arg\max_{\{j \mid t_j < t,\ m(j)=m\}} t_j \tag{8'}$$

With $d_{i,m}^{K,L}$ as the number of individuals at m that transite from K to L at transition time t_i.

The cluster specific baseline hazards $\theta_{0,m.}^{K,L}(t)$ are a compound effect of duration dependence and unobserved workplace specific effects. For our purposes, where we wish to detect whether the observed clustering of absenteeism records across schools are the result of

---

[1] (2') is rearranged at failure points of all ordered durations in the sample. Analogously, the likelihood could be rearranged at failure points of ordered durations within each school/cluster.

sorting or of unobserved workplace effects, it may be convenient to obtain a single measure for unobserved cluster effects. In order to disentangle unobserved heterogeneity from duration dependence, more structure needs to be imposed. It is natural to take the commonly used assumption in duration analysis of time constant unobserved heterogeneity, i.e. as in subsection 2.2 we take equation (3). Next, conditional on the structure of (3), concentrated likelihood methods could be applied on a rearranged version of equation (2') to obtain expression (7) for the unobserved cluster effect $\eta_m^{K,L}$ and (8) for the non-parametric baseline hazard $\theta_0^{K,L}$ (t). As concentrated likelihood and stratified partial likelihood produce identical results, estimates of $\beta^{K,L}$ via (9) could be used to calculate the unobserved fixed effects and the non-parametric baseline hazard. It is important to note that estimates of $\beta^{K,L}$ of the models in subsections 2.1 and 2.2 depend upon the structure of (3), whereas estimates of (9) do not.

In subsection 4.3 we use estimates of $\beta^{K,L}$ of the stratified partial likelihood (9) to recover unobserved fixed effects and perform additional analyses on these. In that section we also make a quick reference to the non-parametric baseline hazard.

In case (9) is viewed as a concentrated likelihood, consistency of estimates relies on asymptotics in time or the number of individuals, taking the number of schools as fixed. This is most approriate for the application in this paper, where we apply the models to absenteeism data of Dutch primary school teachers. In the past decades, the subsequent reductions in the budget of the Dutch education sector has lead to a substantive reduction in the number of schools (by means of mergers) and teachers. Consequently, additional information on absenteeism is expected to come from increasing information over time. If (9) is interpreted as a partial likelihood, consistency of the partial likelihood estimates is also obtained if asymptotics relies on M ($M \rightarrow \infty$). We refer to Ridder & Tunali (1989) for a proof of this.

In order to test the models discussed above against one another we can use Likelihood Ratio tests and Hausman tests. When we compare two parametric models a Likelihood Ratio test can be used. This applies to the comparison of specifications without duration depedence with specifications with a parametric duration dependence and to the comparison of the maximum likelihood estimates of models without fixed effects to a fixed effects specification

that is estimated by the concentrated likelihood method. When comparing a model against a semiparametric alternative a Likelihood Ratio test cannot be used and we will use Hausman tests. These tests follow from the general idea of comparing an estimator that is consistent under the maintained hypothesis to an estimator that is consistent and efficient under the null, but incosistent when the restrictions are violated. This applies to the cases in which we test a parametric maximum likelihood estimator against the stratified or unstratified partial likelihood estimator, but also to testing unstratified against stratified partial likelihood. This can be seen by considering unstratified partial likelihood as a limited information maximum likelihood estimator. The likelihood of observing spell terminations in the order in which transitions are made in the data, irrespective of the durations, is equal to the unstratified partial likelihood (6), with or without fixed effects. In that setting unstratified partial likelihood is efficient, but becomes inconsistent if stratification is required.

3. DUTCH EDUCATION SECTOR AND SAMPLE USED IN THE ANALYSIS

The Dutch Education Sector

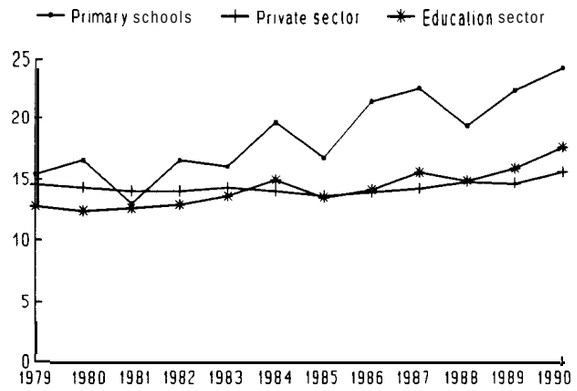
On average about 3.5 million people, i.e. roughly 25 percent of the Dutch population, are participating in the Dutch education system. The sector employs 250,000 workers in 25,000 schools and institutions. One third of the system is managed by (local) government, whereas the remaining two thirds are run by non-profit, denominational, foundations. The system is largely publicly financed. Education expenditures amount to 8 percent of GNP, and account for 20 percent of the government budget.

The education sector in the Netherlands lacks a dynamic sickpay scheme, such as for instance in the U.K. (see Barmby, Orme & Treble (1991a, 1991b)). The sickness benefit program provides a 100% replacement of earnings lost due to mental or physical inability to perform regular duties. We consider primary school teachers, a group of workers facing a uniform (public sector) collective agreement. They are homogeneous with respect to educational achievement and face common wage schemes that are a simple function of functional level and experience. Promotion possibilities within a school are limited and once

tenured, teachers are extremely difficult to discharge involuntarily. Much of these population characteristics reduce the costs of taking absence by means of sickness, and may explain the high incidence rates and long sickness spells.

Figure 1 depicts time series of average sickness spell length in the private sector, in the total education sector and in primary schools. Average duration in the private sector is fairly constant and varies between 13,6 and 15.5 days. The figure for the total education sector is comparable, though it is slightly increasing over time (ranging from 12.4 in 1980 to 17.5 in 1990). The figure shows a markedly different situation for primary school teachers. Where average sickness duration was comparable to that of the other sectors in 1979, it has steadily increased to 24 days in 1990.

Figure 1 Average sickness spells (days) in the private sector, the education sector and in primary schools



Source: Kroniek van de sociale verzekeringen, SvR, 1993 (Dutch SvR Social Security Bulletin 1993). Ziekte verzuim in het onderwijs 1990/1991, Ministerie van Onderwijs en Wetenschappen, 1992, (Sickness Absenteeism in the Education Sector, Ministry of Education, 1992).

In their report on prevention of sickness absenteeism and disablity in the public education sector, the Ministry of Education (1992) notes that sickness absenteeism in this sector is not only-much higher than in other sectors, but also that absenteeism is highly concentrated among a relatively small group of workers. In 1989 only 13.5% of the employees in the education sector accounted for a total of 80% of all days lost due to sickness absenteeism. Furthermore, it was noted that mental inability to perform regular duties was one of the major

causes for Disability Insurance entrance. It is suggested that teachers are more frequently and more persistently exposed to stressful situations than their counterparts in the private sector. This is partly due to difficulties with teaching itself (caused for instance by size of classes and pupils' attitude towards school), but more to problems encountered in the work environment. Relational problems with colleagues, work ethic within the school, school's management, and limited promotion possibilities for teachers are considered to be the major determinants.

In the mid eighties, by means of an experiment, for some schools schools' health services, previously provided at the regional level, were organized at the school level. These school health services had to provide medical as well as psychological and social assistance to school employees. Moreover, the staff of the school health services included a specialist in the field of organization of firms and firms' labour management to support school's management. By providing these services at a local, decentralized, level it was thought that most of the major causes of absenteeism (listed above) could be neutralized. Some schools in our sample were included in this experiment. Therefore part of the discussion in section 4 will be devoted to the effect of health services on sickness absenteeism.

Data

The data consist of sickness absenteeism records of education sector workers registered by the Leiden Institute for Social Science Research on behalf of the Ministry of Education. The total sample consists of about 30000 unique employees and 1100 schools (primary, secondary and higher education) that have been surveyed for on average 3 years over the period 1987 to 1991. From this sample we select schools at the primary level resulting in a set of 426 schools consisting of 4969 teachers accounting for 21137 spells of sickness and work.

All employees within a school are observed from the moment their employer enters the sample, or, from the moment they start working at a school that is already participating in the survey. Analogously, individual observations stop either when the school leaves the sample or when staff leaves the school. In the latter case the exact destination is often unknown. For that reason we abstract in our model from differences between alternative exit routes out of the job. Implicitly it is assumed that all these categories can be lumped together into a single

job exit category. We are primarily concerned with sickness absenteeism behaviour, and therefore concentrate the analyses on two dimensions of sickness absenteeism: sickness incidence (associated with the transition from work to sickness) and sickness duration (associated with the transition from sickness to work). The tables below give a first impression of both dimensions of sickness absenteeism in our data.

From tables la and lb we can see that 21137 spells are observed in total, of which 12836 are work spells and 8301 sickness spells. The majority of the sickness spells is of a short term nature, 82.9% of the observed sickness spells does not exceed 14 days. On the other hand a substansive number of sickness spells may last for several moths, or may even exceed a year. As a result mean sickness spell length approaches four weeks (27.26 days).

Table la Cross tabulation of spells

	Work	Sickness	Exit	Censored	Total
Work		8097	527	4153	12836
Sickness	7923	-	78	300	8301
Total	7923	8097	605	4453	21137

Table lb Distribution of sickness spells in the sample

Length (days)	# spells	cumulative percentage
1		20.4
[2,3]	1692 1717	41.4
[4,7]	2562	72.0
[8,14]	908	82.9
[15,42]	640	90.6
[43,182]	474	96.3
[**183,365]**	193	98.6
[365,→)	115	100.0
Mean spell length	27.26	

Aggregate aggregate measures of sickness absenteeism in our data reveal that the distribution of average spell1 length per individual and school are heavily skewed to the right. At the individual level 75% of the teachers with at least one spell experience average spell length of

two weeks or less. Though the majority of the individuals experience relatively short spells, their share in total number of days lost to absenteeism is limited. These individuals account only for 15% of the total number of days lost due to absenteeism (i.e. those with average spell lengths exceeding two weeks (only 16% of the total sample) account for 85% of the total number of days lost due to absenteeism). The fraction of schools with average spell length not exceeding two weeks is 44%. These schools (197) account for only 15% of total days lost to sickness absenteeism. Schools with average spell length exceeding 50 days (83 out of 426 schools) account for 62% of total days lost to sickness absenteeism. We may therefore conclude that absenteeism is concentrated among a relatively small number of schools. These numbers are in line with previously mentioned results of the Ministry of Education (1992).

Although evidence of this type is often used to suggest differences between schools, it is difficult, if at all possible to find direct support for the clustering hypothesis in the numbers presented above. The distribution of average durations per school may reflect the uneven distribution of sickness spells over teachers. Moreover, the schools differ with respect to the number of teachers employed. The shape of the distribution of average sickness spells over schools may therefore be a perfectly ordinary statistical phenomenon: some schools having the bad luck to have hired sickness-prone teachers, other schools being more lucky. In order to test whether clustering is present or that the observed distribution is a result of a fair lottery, we performed a non-parametric (Kolmogorov-Smirnov) test. The test supports the clustering hypothesis and is documented in a companion paper (Lindeboom & Kerkhofs (1995)).

The question whether clustering is caused by circumstances specific to the workplace or by a sorting of teachers, can only be adressed by models that allow for observed and unobserved differences between teachers and schools/workplace in a flexible way. As far as observed characteristics are concerned, our dataset contains a variety of personal characteristics and school (environmental) characteristics. Table Al presents means of the main variables used in the empirical analysis. We postpone a discussion of these variables to subsection 4.1.

4. EMPIRICAL IMPLEMENTATION AND RESULTS

4.1. Empirical Implementation

Before we turn to the results, we first discuss the empirical implementation of the models presented in section 2. This concerns a discussion of sampling issues, specification, identifiability of the unobserved group effects and the treatment of time varying covariates.

Sampling issues

We sample 4969 teachers from 426 schools at a specific point in time and follow them subsequently until they either leave the school or the school leaves the sample. Since censoring is observed as the school leaves the sample, we have to assume that the time at which a school leaves the sample is of no influence for the individual hazard rates of teachers within the school. This assumption guarantees that, if censored, individual failure times (durations) are independently right censored. This (harmless) assumption, is very convenient since it protects us from modelling the censoring mechanism jointly with the individual failure times.

Given this sample set up, a likelihood function can be constructed that consists of the product of stock sampled first spells and subsequent flow sampled spells. Explicit expressions for the stock sampled spells are given in Flinn & Heckman (1982), Ridder (1984) and Lancaster (1990). In general, stock sampled spells require joint modelling of the probability of entrance in the first observed state. To quote Lancaster (1990, pp189): '. . . we require to imbed the stock sampled data in a stochastic process describing the full bibliography of each individual.. . '. This implies that in general the proportionality of the hazard of the stock sampled durations is lost, making the non-parametric model of section 2 (virtually) non-estimable. Flinn & Heckman (1982) propose to specify a seperate duration distribution for first sampled spells (see e.g. Gritz (1993) and Ham & Lalonde (1996) for applications). In either case, solutions for the initial condition problems require information on the elapsed duration at the sampling date, unless absence of duration dependence is assumed. As we do not observe elapsed duration in a work spell at the date of initial selection, and observe stock

sampled sickness spells with error[2], we proceed in a different way by conditioning on the first sampled spells. Under the assumption that all individual variation can be described by x and $\eta_m^{K,L}$ (cf section 2) a likelihood can be constructed that omits the first (stock) sampled spells and consists of the product of the remaining flow sampled spells. So, basically we construct a sample from the initial sample set up by following individuals over time, and restrict the attention to newly started spells after the initial selection date. This results in simple likelihood expression as in section 2[3].

Specification

The unit of time in our empirical models is taken to be one day. The part of the transition rates associated with the observed regressors are specified as exponential functions. For instance, we specify $\exp\{x'\beta_1\}$ for the transition form work to sickness, $\exp\{x'\beta_2\}$ for a work to exit transition etc. The set of regressors x include individual characteristics and school characteristics. Most of the included variables are time varying in the sense that they are allowed to change in the beginning of each school year. However, regressors are taken as fixed during the course of a specific spell.

Included individual characteristics are Age (measured in years), Gender[4] (dummy for females), Married (dummy for those living together with partner or married), Permanent contract, Tenure1 (linear job duration effect for jobs lasting less than 5 years), Tenure2 (linear job duration effect for jobs lasting longer than 5 years), Part-timer (dummy variable

---

[2] Schools are sampled at the beginning of the school year. The exact length of the sickness spells that start during the summer vacation preceding the school year is not known. Moreover, stock sampled sickness spells that started prior to the summer vacation are suspect of being recorded with error.

[3] Note that the fixed effect approach simplifies the likelihood considerably. In a random effect approach additional assumptions are required in order to obtain tractable expressions for the likelihood. The terms $\eta_m^{K,L}$ need to be independent of the included regressors x, and we need the assumption that the unobserved components of the alternative states are independent of each other. This last assumption is effectively the semi-Markov assumption.

[4] Pregnancy leave may distort the distribution of spell incidence and duration, as in the Netherlands the pregnancy leave period is statutory fixed at 16 weeks. The files do not allow for a distinction between different causes of sickness absenteeism. We therefore spotted the data to reveal whether there was a clustering of sickness spells around 16 weeks. If pregnancy leave had a significant effect on the duration distribution of sickness spells, one would expect to see this in the data. We did not find any evidence for this. We refer to our companion paper (Lindeboom & Kerkhofs (1995)) for more details.

for part-time workers), Small groups (dummy for those teaching at small (≤ 20 pupils) classes), Large groups (dummy for those teaching at large (≥ 31 pupils) classes), Lower groups (dummy for infant school teachers) and Head (dummy for head of the school).

The set of school variables is included to capture, workplace effects, schools' ability to replace absent workers, and school's management towards absence behaviour and the presence of a Health service at the school (previously discussed in section ***3).*** The variables Short replacement easy/difficult, and Short sickness easy/difficult are included to capture the school's ability to replace teachers for a short period, and/or the school's ability to cope with short term sickness absenteeism. Difficulties in replacing (absent) workers may induce additional stress on fellow teachers who need to replace these. This may lead to further absenteeism. It may be clear that these variables could be used as key instruments in policies to fight sickness absenteeism.

In the mid-eighties the government attempted to reduce its budget by means of mergers between schools within the same region. Our sample includes some of these schools. Clearly, mergers are associated with changes in workplace situations such as change of school board/management, number of teachers, number of pupils per group etc. The variable Merger is included to capture the effect of this. Variables Catholic, Protestant, Urban, Rural, Number of teachers and Pupil size decreasing/increasing are included for obvious reasons. The same holds for the Avge variables. These are school average variables for the fraction of females, the average age, the fraction of teachers teaching at lower groups (infant school) etc.

The Catholic variable needs some special attention. For a few schools in our sample, in the course of the years that we follow them, the denomination changes from Catholic to the reference category (Public or Special). Presumably, this change in denomination is caused by a change of the school board, school's management and/or a merger in the sample period. Unfortunately our dataset does not provide this information. It should therefore be noted that Catholic may capture more than a pure effect of denomination.

Identification

Both school specific fixed effects as well as variables at the school level are identified by

accross school variation. As a consequence, to identify school specific fixed effects from school variables, sufficient independent variation over time of the (set of) school variables is required. Therefore, in the estimation of the fixed effects models time constant variables at the school level, such as the Avge variables, Protestant, Urban, Rural and Merger are not identified. Effectively, they are absorbed by the unobserved school specific fixed effects. We will relate the fixed effects to the time constant school variables in the analyses of section 4.3.

***Models** estimated*

We estimated all models of section 2. We also estimated partial likelihood and maximum likelihood models that do not allow for unobserved workplace effects. Recall that schools with no relevant transition do not contribute to the likelihood for models that account for unobserved workplace effects. Therefore, effectively, these models are estimated on a smaller sample than more traditional model that do not allow for unobserved cluster effects. Consistency of the parameters β of the fixed effects models is guaranteed under the model assumptions, though the estimates may be affected in small samples. We expect little effects from this sample requirement. For the transition from Sickness to Work only 2 spells (out of 8094) were omitted for the estimation of the fixed effect models. For the transition from Work to Sickness 68 spells (out of 8251) were omitted.

Tables 2a and 2b report estimates of models that account for school specific fixed effects, using concentrated and stratified partial likelihood methods. Columns one and two of each table report results of a model with time constant fixed effect that are concentrated out of the likelihood (specification (5) of section 2). Regression parameters β are estimated along with the parameters of the baseline hazards. The first column reports on results for a model without duration dependence, denoted as specification I. The results for models with a limited set of duration dummies are reported in column two. We denote these as specification II. Column three presents the results from a the most flexible model in which baseline hazard and fixed effects are left unspecified. The regression coefficients are estimated using the stratified partial likelihood (9) of section 2. We denote these as specification III in the tables. We also estimated the partial likelihood model of section 2.2 that allowed for time constant

unobserved workplace effects. The estimates of these models were virtually identical to those of specification II. Specification II is more appealing as it relatively straightforward to estimate. We therefore present these instead of the results of the model of section **2.2.**

Conditional on estimates of specification III, we use (7) and (8) to recover the non-parametric duration dependence function and the fixed effects. More details on this procedure are provided in subsection 4.3, where we also discuss results from additional analyses on the fixed effects. The non-parametric duration dependence functions are depicted in figures Al and A2. From these figures it can be seen that both the sickness incidence and sickness duration display strong negative duration dependence. This picture is most pronounced for the transition from Sickness to Work, where the hazard rate falls sharply after the first few days.

A comparison of results for all these models may give an indication as to what extend correcting for school specific fixed effects in a flexible way is important for the parameters of interest (β). In a companion paper (Lindeboom & Kerkhofs (1995)) we also present estimates of traditional models that do not allow for unobserved workplace effects. We briefly report on this when we discuss the results of specification I, II and III. The traditional models, the model of section 2.2 and specifications I, II and III are more formally compared at the end of subsection 4.2. The tests indicate that there is strong evidence in favour of stratification. We therefore mainly concentrate on the results of the most flexible, stratified partial likelihood, model (specification III).

4.2 Results

The transition from work to sickness ($W \rightarrow S$)

We start with the results from the most flexible model, specification III. Both individual characteristics and school characteristics are of importance for sickness incidence, though individual characteristics appear to dominate in size. With respect to individual characteristics we find strong positive significant effects of the variables Female and Permanent contract. As documented in section 2, tenured teachers are extremely difficult to discharge involuntary, and in case of sickness a 100% replacement of earnings is provided. As a consequence,

claiming is virtually costless for tenured teachers, which might induce them to report sick more often than their counterparts with no permanent contract.

Significant negative effects are found for Part-time workers and those teaching for small groups/classes. The age effect, modelled by a quadratic function, is maximized at age 41. This implies for instance equal age effects for a 21 year old teacher and an elderly teacher of 61. This might be explained by what may be called a survivor effect. As in most other OECD countries participation rates of Dutch elderly workers has declined dramatically in the past decades. This is particularly true for public sector education workers. The share of older workers (55 and over) in this sector amounts to only 5%. The bulk of the teachers either retires or changes profession considerably before the mandatory retirement age. It may that the few teachers that remain working and retire later are more committed to their profession.

< Table 2a around here >

Only two main school variables appear to be of interest. Spell incidence is on average higher for teachers at schools that have difficulties in replacing teachers for a short time period. The opposite effect is found in case schools have difficulties in replacing absent (due to sickness) workers. The difference in these two variables is mainly that the replacement variable is associated with an anticipated need for replacement, whereas the sickness variable is associated with a sudden, unanticipated need for replacement of absent workers. It is conceivable that unanticipated additional work (for teachers) may induce them to postpone sickness absenteeism. If so, however, one would expect sickness spell duration to increase. A check on the results for the $S \rightarrow W$ (Table 2b) transition reveals that such an effect is present. The effect, however, is not significant at the standard levels. With respect to the remaining school variables it is interesting to note that little effect of school Health services is found.

A comparison of the estimates of specification I and II on the one hand and specification III en the other hand shows that allowing for more flexibility both duration dependence and unobserved school effects has little effect on the parameter estimates of the individual variables, but that it has some effect on the parameter estimates of the school variables. In absolute value almost all of these parameter estimates reduce in size. A particularly interesting

variable in this respect is the variable Health service. The significant effect of this variable in the model with no duration dependence appears to be spurious.

In order to evaluate the importance of school specific effects for the regression parameters (β), we also compared the results of Table 2a with the results of models that do not allow for unobserved workplace effects. It can be concluded that mainly the parameter estimates of the school variables are affected, in case one does not allow for school specific fixed effects.

The transition from sickness to work ($S \rightarrow W$)

From columns three Table 2b we can see that in general the coefficients of both individual and school characteristics are relatively small as compared to the coefficients associated with sickness incidence reported in Table 2a. Gender, marital status, whether one has a permanent contract, and pupil class/group size are individual (teacher) characteristics that have a significant effect on sickness duration. Most signs of these parameter estimates are as expected. For instance females or those with a permanent contract experience longer sickness absenteeism spells etc. The non linear effect of class/group size is a little more puzzling. This could be explained by the fact that at most schools, more able and more experienced teachers are assigned to groups with a larger number of pupils. The effect of Age is negative over the relevant range, implying that elderly have on average longer sickness absenteeism spells.

With exception of the Health services variable, surprisingly little effects are found from the school variables on the exit rate out of sickness. Hence, conditional on the unobserved school specific fixed effect, remaining variation of sickness absenteeism duration seems to come mainly from variation in individual characteristics. Of course, the relative importance of the school specific effects in explaining total variation still remains to be assessed. We do this in section 4.3. Schools with health services have on average shorter sickness absenteeism spells. As discussed in section 3, these health services were introduced by the government by means of an experiment in order to fight sickness absenteeism at the school level. It has to be noted however, that though significant at the 5% level, the size of the effect of health services on sickness duration seems to be moderate.

< Table 2b around here >

It is important to allow for duration dependence in modelling sickness absenteeism duration. From a comparison of specification I with II, one can see that strong duration effects are found. An initially slightly increasing exit rate displays strong negative duration dependence afterwards. Normalizing the baseline situation as one, the probability of returning to work (per day) reduces to only $\exp\{-3.16\} = 0.04$ after 42 days, making sickness a virtually absorbing state. The relevance of duration dependence for the regression coefficients can also be seen from a comparison of the columns of Table 2b. Both the parameter estimates of individual characteristics as well as those from school characteristics change considerably. A comparison of specifications III and II also reveals that though most parameters estimates remain stable, some notable changes occur for the $S \rightarrow W$ transition rate. The effect on the school health services variable is discussed above, but there are also notable effects on parameter estimates of individual characteristics. The effect of Permanent contract is reduced, whereas the variables Head, Small groups and Health services gain in size and significance.

We found from a comparison of specifications I, II and III and models without unobserved fixed effects that in modelling sickness absence duration it is important to allow for duration dependence and school specific fixed effects in a flexible way in order to avoid biases in the parameter estimates and the conclusions that can be drawn from these.

Transitions out of the job

In section 3 we argued that estimation of a school specific effect (η) requires at least one relevant transition, in a school m. This issue becomes particularly relevant for the $(S \rightarrow Exit)$ transition. Estimation of this exit rate appeared to be impossible due to the limited number of transitions of this type (Only 78 $S \rightarrow Exit$ transitions are observed, see Table la). For this reason we only report fixed effects estimates for the $W \rightarrow Exit$ transition. These are reported in Table A2 of the Appendix. Below we give a brief discussion of the main results.

Individual characteristics are of more importance than observed school characteristics in explaining job exit behaviour. Duration dependence seems to have little effect on the parameter estimates and is found to be consistent with predictions from existing job turnover models (Initially increasing exit rates fall as time proceeds). Informal comparison of estimates

of this model with those of duration models without fixed effects indicate that notably the parameter estimates of the school variables are sensitive to the inclusion of unobserved group effects.

Comparison of alternative models

In section 2 we discussed a range of alternative models that could be estimated to test for the relevance of unobserved school/cluster effects and/or the importance of duration dependence. Table 3 below summarizes the findings.

Table 3 Testing of alternative models

	Work to Sick	*Sick to Work*
Testinp for duration dependence (no fixed effects)		
no vs. parametric (5 steps)	LR*=281.96 (9.49)	LR=10211.3 (9.49)
parametric vs. unstratified PL	TH=40.14 (27.6)	TH= 1002.2 (26.3)
unstratified vs. stratified PL	H=74.56 (32.7)	H=62.20 (32.7)
Testing for duration dependence (with fixed effects)		
no vs. parametric (5 steps)	LR= 104.44 (9.49)	LR=7302.6 (9.49)
parametric vs. unstratified PL	TH=345.2 (19.7)	TH=1277.9 (26.3)
unstratified vs. stratified PL	H=81.15 (32.7)	H=239.8 (32.7)
Testing for school specific fixed effects		
no duration dependence	LR=1186.0 (474.1)	LR=3515.0 (472.0)
5-step duration dependence	LR=1008.5 (474.1)	LR=606.3 (472.0)
non-parametric duration dep.	TH=50.68 (31.4)	H=69.16 (32.7)

5% critical values between parentheses.
*** LR=Likelihood Ratio test; H=Hausman test; TH=Hausman test excluding negative eigenvalues of the covariance matrix.**

All tests reject their null hypothesis. This strongly underlines the importance of duration dependence, even if unobserved school specific differences are taken into account. The

importance of accounting for these differences by introducing fixed effects into the specification is also firmly supported by these tests. However, it has to be added that the fixed effect estimates with **5-step** duration dependence, fixed effect unstratified partial likelihood and stratified partial likelihood lead to estimates that are very similar. In these cases the Hausman test weighs the fact the estimates are hardly affected by imposing the restrictions against the resulting efficiency gain. The latter is even smaller and the test concludes that the restrictions should not be imposed. One could argue that the effect of the restrictions on the parameters of interest is negligable and therefore the more restrictive specification, the fixed effect model with **5-step** duration dependence is preferred.

The fact that some estimates were almost identical lead to the numerical problem that the difference between the covariance matrices is not positive definite. In that case we have used a truncated version of the Hausman test (TH in table 3). The difference between the covariance matrix is written as a diagonal matrix of eigenvalues that is pre and post-multiplied by a matrix of orthonormal eigenvectors. When computing the inverse of that difference matrix we use the reciprocal of an eigenvalue only if it is positive, using 0 otherwise. The number of degrees of freedom of the test is equal to the number of positive eigenvalues.

4.3 School specific fixed effects reconsidered

The previous subsections were concerned with the effect of unobserved school specific effects on the regression coefficients. What remains to be answered is the relative importance of unobserved school effects on sickness incidence and sickness duration. Moreover, it still remains unclear whether the unobserved school specific effects can account for the large variation in sickness absenteeism behaviour across schools, and the apparent clustering of schools with short sickness absenteeism records (‘healthy’ schools) and those with long sickness absenteeism records (‘sick’ schools). These issues were previously noted in a report of the Ministry of education (see also section 3). As the prime goal of this section is concerned with sickness incidence and sickness duration, we omit results of job exit behaviour .

We use the results from the most flexible model to tackle these remaining questions. As described in section 3, this model accounts for school specific non-parametric baseline hazards. In a way these baseline hazards are a compound effect of duration dependence and unobserved heterogeneity. We take the commonly used assumption that unobserved school effects are constant over time (cf section 2, equation (3)), in order to disentangle duration dependence from unobserved heterogeneity. Note that this structure is imposed after that we have estimated the stratified partial likelihood (9). As a consequence, violation of this assumption is of no influence for estimates of β. Given estimates of β from (9), time constant unobserved school specific effects can be calculated from the non-linear system of 426 equations with 426 unknown school effects (equation (7)). The unobserved school effects are identified up to a scale factor. Next (8) can be used to solve for the non-parameteric baseline hazards. These baseline hazards are common to all individuals in the sample and could be interpreted as Kaplan-Meier estimates, after a proper reweighting of the data with β and the η. Figure Al and A2 of the appendix present the baseline hazards for sickness incidence and sickness duration. We now turn to the analyses presented below.

The relative importance of school specific effects

Table 4 is included to assess the relative importance of fixed effects in explaining absenteeism behaviour across schools. The table reports the school averages of the regression part $\exp\{x'\beta\}$ and the fixed effect η of the exit rates of $W \rightarrow S$ and $S \rightarrow W$. As noted above, school specific effects are identified up to a scalefactor. We normalize the mean of the fixed effects $\exp\{x'\beta\}$ to one.

As far as sickness incidence is concerned, Table 4 reveals that the variance of η relatively large as compared to the variance of $\exp\{x'\beta\}$. Furthermore, judging from the third and fourth order moments of the distributions, the distribution of fixed effects is more heavily skewed to the right and has fatter tails than the distribution of $\exp\{x'\beta\}$. This picture is even more pronounced for the $S \rightarrow W$ transition. The regression function $\exp\{x'\beta\}$ hardly varies, and is approximately symmetrical. On the other hand the distribution of the fixed effects is characterized by a relatively large variance, a large skewness parameter and has fat tails. We

may conclude from this that the dispersion in school specific fixed effects dominate that of the observed (teacher) characteristics, and resembles the observed variation in sickness absenteeism accross schools.

Table 4. Distribution of school specific fixed effects and exogenous variables in the sample'

	mean	st. dev.	skewness	kurtosis
Fixed effects				
$W \rightarrow S$	1	0.644	0.963	2.448
$S \rightarrow W$	1	0.426	1.260	2.865
regression function $\exp\{x'\beta\}$				
$W \rightarrow S$	1	0.174	0.101	0.683
$S \rightarrow W$	1	0.103	-0.046	0.467
$\text{corr}[\eta^{S,W}, \eta^{W,S}] = -0.15$				

[1] Statistics are derived for the sample of 426 schools. For each school, $\exp\{x'\beta\}$ and $\exp\{x'\gamma\}$ are unweighed within school averages.

Table 4 is informative in the sense that it can tell us something about the relative importance of school specific effects as compared to the regression functions, and it also enables us to see whether dispersion of the fixed effect or that of the regression functions may account for observed dispersion of sickness absenteeism behaviour. The table can not tell us whether observed sickness incidence or duration within a specific school, is a result of a large or small school specific effect or of the composition of exogenous characteristics within the school (the sorting effect). We use Figure 2 to see which of the two effects is dominant in our data.

Figure 2 consists of four parts. In part 2a and 2b we confront school specific effects (2a) and regression functions (2b) of the $W \rightarrow S$ transition with observed sickness incidence records in our sample. Similarly, part 2c and 2d are scatter diagrams of schools' unobserved fixed effect and observed mean sickness duration, and of schools' regression function with observed mean sickness duration, respectively. A sorting effect is present if one can find a positive (negative) association between the effect of the exogenous variables $\exp\{x'\beta\}$ on sickness

Figure 4c

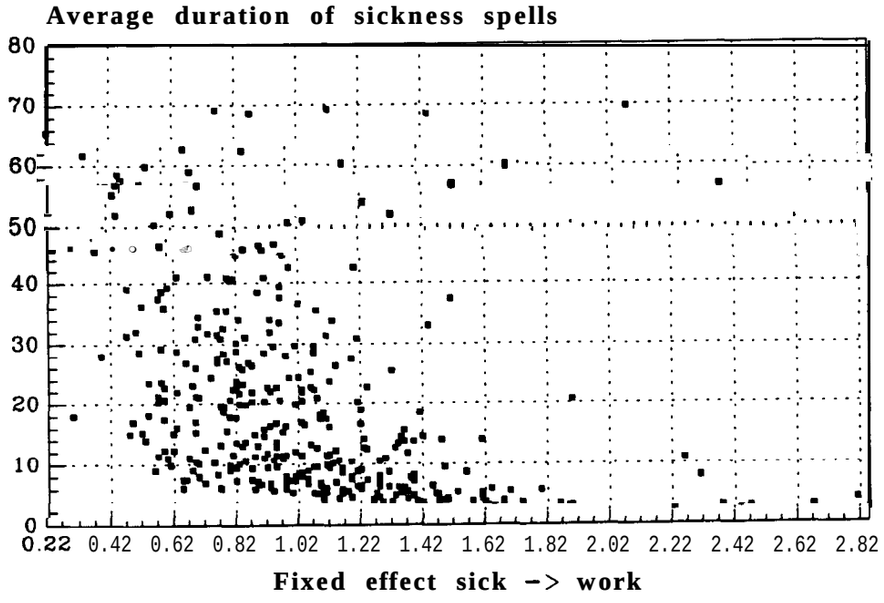
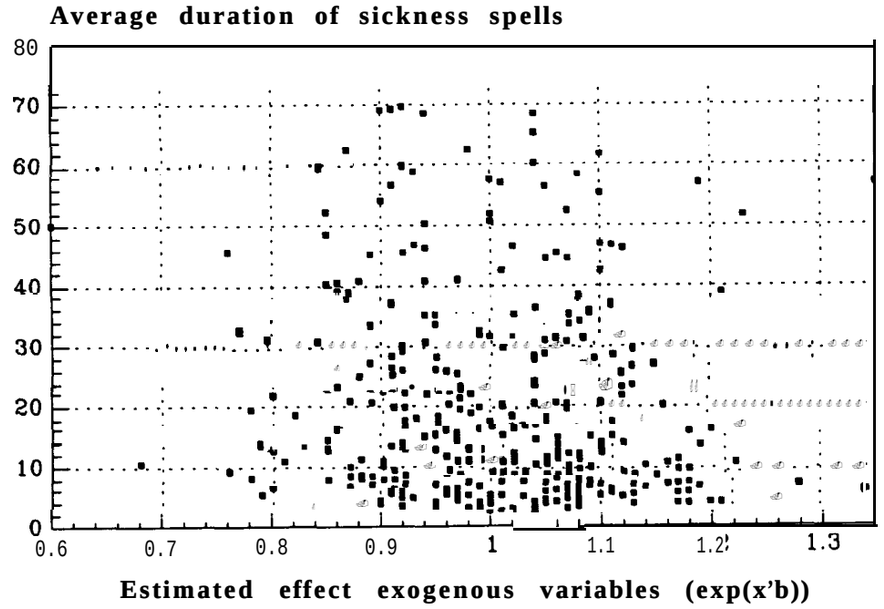


Figure 4d



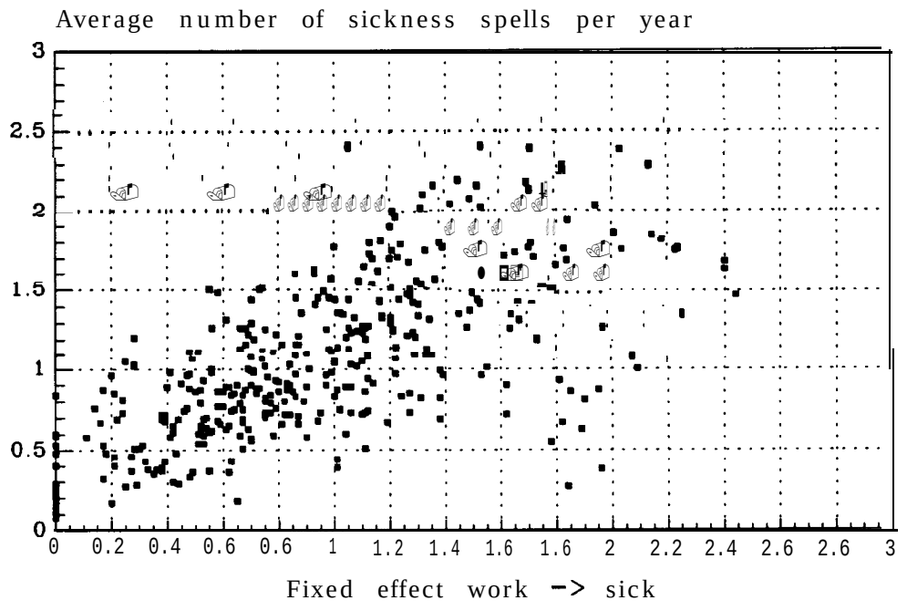
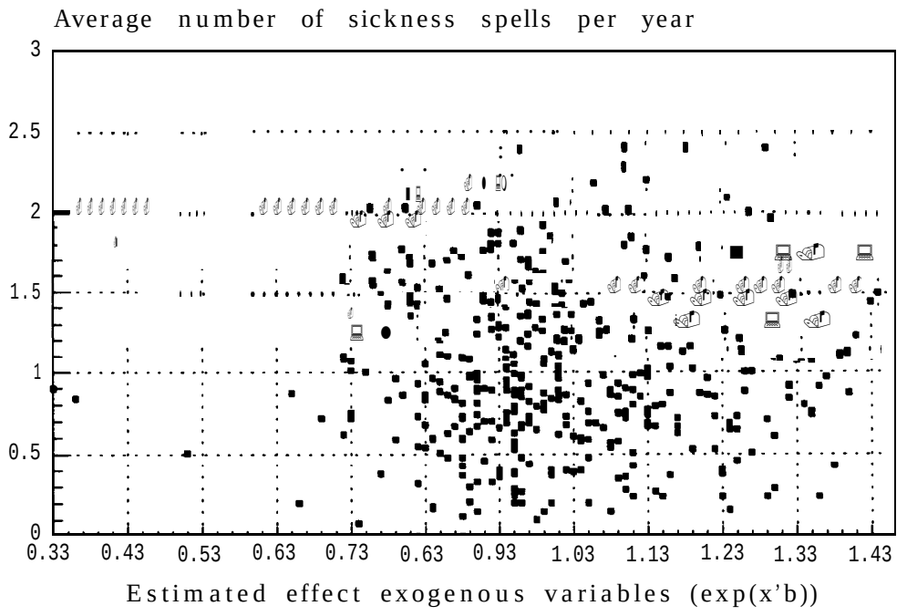

Average number of sickness spells per year
3
2.5
2
1.5
1
0.5
0
0 0.2 0.4 0.6 0.6 1 1.2 1.4 1.6 1.6 2 2.2 2.4 2.6 2.6 3
Fixed effect work -> sick

Figure 4b


Average number of sickness spells per year
3
2.5
2
1.5
1
0.5
0
0.33 0.43 0.53 0.63 0.73 0.63 0.93 1.03 1.13 1.23 1.33 1.43
Estimated effect exogenous variables (exp(x'b))

incidence (duration) in figure 2b (2d). In case unobserved school effects dominate the observed patterns of sickness absenteeism one should detect this in figures 2a and 2c. From Figure 2a it can be seen that high school incidence records are associated with on average larger values of the school specific effects. This picture is apparently not present in figure 2b. There does not seem to be a specific pattern between sickness incidence scores per school and within school averages of the regression function $\exp\{x'\beta\}$. Within each class schools exists with relatively high and relatively low values of $\exp\{x'\beta\}$. The correlation between $\exp\{x'\beta\}$ and sickness incidence is -0.0909 and insignificant at the 5% level. Hence any clustering in sickness incidence records among specific schools, can not be ascribed to a clustering of individuals with 'bad' characteristics to schools with high records, and those with 'good' characteristics with schools with low incidence records. A 'sorting' effect seems to be absent in our data. Instead, it appears from figure 2a that any clustering in the data could be ascribed to the fact that schools with low incidence rates ('healthy' schools) have on average lower unobserved school specific effects. We find a strong significant correlation of 0.7198.

Figures 2c and 2d display a similar pattern. Though less prominent as in figure 2a, there appears to be a inverse relationship between $\eta^{S,W}$ and observed sickness duration, i.e. schools with short average durations experience on average larger values of η. The correlation between $\eta^{S,W}$ and sickness duration is -0.2965 and is significant at the 5% level. Again, the regression function $\exp\{x'\beta\}$ does not seem to be related to observed average duration in a school (the correlation, -0.0904, is insignificant). As a consequence, observed clustering in the data is more likely to be a result of a school environmental effect.

An analysis of school specific effects

The fixed effects that are found to be important in the previous section support the hypothesis that the clustering of schools is caused by differences between schools. In the estimates in section 4.2, we had to omit all constant exogenous variables referring to school characteristics, in order to identify the school specific effects. Effectively, the effect of these control variables are encompassed by the school-specific effects. In this section we relate the school specific effects to the variables characterizing the school environment. This analysis

serves two purposes. First of all to explain as much as we can why schools are different and secondly to see to what extent the differences can not be related to directly observable characteristics of the type we have in our data. Table 5 contains the results of a simple regression of the estimated fixed effects per school on exogenous variables characterizing its size, denomination, composition of the teaching staff, short term replacement opportunities in case of expected and unexpected absenteeism and the presence of health services. Seven schools had to be omitted because of missing observations on right hand side variables. The estimated fixed effects of the partial likelihood estimates are used and for the transition from sickness to work, 7 outliers - for schools with a small number of extremely short sickness spells - were omitted.

Table 5 Least squares estimates of the fixed effects on school characteristics[1,2]

	Work -> Sick		Sick -> Work	
constant	0.5829	**(1.00)**	1.5919	(3.86)
number of teachers	0.0191	(1.94)	-0.0033	(0.77)
number of pupils	0.0010	**(1.64)**	0.0003	(0.77)
proportion of lower grade groups	0.0484	(0.85)	0.0201	(0.46)
average number of pupils in group	-0.0094	(0.94)	-0.0116	(1.42)
average age teachers	0.0038	(0.41)	0.0050	(0.72)
number of females	-0.0276	(0.15)	-0.0279	(0.24)
number married	0.2259	**(1.40)**	0.0283	(0.22)
number tenured	0.2381	(0.53)	-0.4788	(1.69)
average job tenure	-0.0109	(0.93)	-0.0063	(0.84)
catholic school	-0.4320	(5.59)	-0.0378	(0.84)
protestant school	-0.2560	(3.25)	0.0069	(0.14)
school has merged	0.5214	(1.26)	-0.2967	(1.93)
merger expected	-0.1516	(0.67)	0.2054	(2.05)
health service present	0.1377	(2.14)	0.0278	(0.68)
short term replacement				
anticipated replacement easy	-0.1725	(2.50)	-0.0677	(1.32)
anticipated replacement hard	-0.2411	(2.49)	0.1259	**(1.64)**
unexpected replacement easy	0.0607	(0.89)	-0.0349	(0.66)
unexpected replacement hard	0.4469	(3.67)	-0.0098	(0.16)
Adjusted R-squared	0.195		0.0136	
F-statistic	6.23		1.28	
# observations	419		410	

[1] Absolute t-values in parentheses, based on White's heteroscedasticity consistent covariance matrix of the estimator.

[2] Some of the school variables are time varying. We take their value at the date of selection.

From the estimates it follows that sickness incidence is higher on large schools where a health service is present and significantly lower on catholic and protestant schools. The effects of replacement opportunities may exhibit an endogeneity problem. Whereas it may be expected that sickness incidence is lower if replacement is hard to arrange, schools that have high incidence rates will typically find it more difficult to arrange replacement for sick teachers. With respect to sickness durations, the only significant variable is the number of tenured teachers. Schools with a low proportion of tenured teachers (typically younger teachers) show significantly shorter sickness durations. For policy purposes it is important to notice that the presence of a health service does not significantly reduce the average sickness duration, but significantly signals a high incidence rate. Apparently the health services are only partially succesful in reducing sickness incidence. Most importantly, the estimates indicate that the school specific effects are hardly related to the exogenous variables of the type available in our data. The coefficients of determination are low, as are the F-statistics. Although it is clear that school-specific conditions affect sickness absenteeism records, further research into the idiosyncracies of sick and healthy schools are called for.

6. CONCLUSIONS

In the Netherlands sickness absenteeism of public school teachers is known to be notoriously high, to vary considerably among schools and there appears to be a clustering of absence data. We focus on sickness incidence and sickness duration of individual teachers within a school to assess whether sorting effects or workplace characteristics cause the large variance and clustering in the data. We specifiy and estimate concentrated and parial likelihood models that allow for unobserved workplace effects. The most flexible model is a stratified partial likelihood model that allows for non-parametric school-specific baseline hazards. We show that this stratified likelihood can be derived using a concentrated likelihood approach. This concentrated likelihood approach allows us to recover estimates of unobserved workplace effects and non-parametric baseline hazards given estimates of the regression coefficients

obtained from the stratified partial likelihood. The unobserved workplace effects are used to detect the causes for the observed variation and clustering in the absenteeism records.

In the analyses we find strong effects of both observed personal characteristics and school - characteristics. From a comparison of a range of models we conclude that it is important to allow for unobserved workplace/school effects, but that this also needs to be done in the most flexible way. Unobserved workplace specific effects account to a large extent for the observed variation of sickness absenteeism across schools. We also find that the observed clustering in 'healthy' schools and 'sick' schools is a result of unobserved school effects instead of a teacher sorting effect. In an additional analysis we relate the school specific fixed effects to a range of observed exogenous school variables. The estimates indicate that the school specific effects are hardly related to the exogenous variables of the type available in the data. It remains however, that workplace effects are important in explaining sickness absence patterns, and a better understanding of these workplace conditions will prove to be essential in reducing sickness absenteeism.

Table 2a. Estimation results of models with school specific fixed effects. The transition from work to sickness *(W → S)*[1]

	I		II		III	
i) Variables at the individual level						
Female	0.16	(4.3)	0.15	(4.0)	0.12	(3.2)
Age/10	0.48	(3.0)	0.42	(2.6)	0.49	(3.0)
(Age/10)^2	-0.06	(3.1)	-0.05	(2.7)	-0.06	(3.1)
Married	-0.06	(1.8)	-0.06	(1.7)	-0.07	(1.9)
Perm . contract	0.29	(4.3)	0.27	(4.1)	0.21	(3.0)
Part-timer	-0.20	(5.3)	-0.19	(5.0)	-0.17	(4.2)
Head	-0.05	(1.1)	-0.05	(1.1)	-0.04	(0.8)
Lower groups	0.09	(2.6)	0.08	(2.4)	0.09	(2.5)
Tenure1	-0.04	(1.0)	-0.03	(0.9)	0.00	(0.0)
Tenure2	0.08	(0.5)	0.08	(0.6)	0.15	(1.0)
Small groups (≤ 20)	-0.10	(2.6)	-0.09	(2.6)	-0.09	(2.5)
Large groups (≥ 3 1)	-0.08	(1.6)	-0.08	(1.6)	-0.07	(1.3)
ii) Variables at the school level						
Catholic	*0.39*	(1.8)	0.29	(1.3)	0.24	(1.1)
# of teachers	-0.01	(1.3)	-0.01	(1.3)	-0.01	(1.3)
Pupil size decreasing	0.05	(0.8)	0.05	(0.8)	0.03	(0.5)
Pupil size increasing	0.13	(1.9)	0.12	(1.8)	0.10	(1.4)
Health services	-0.10	(2.0)	-0.07	(1.5)	-0.08	(1.6)
Short replace. easy	0.13	(1.9)	0.12	(1.7)	0.10	(1.3)
Short replace. diff.	0.31	(3.6)	0.26	(3.1)	0.22	(2.5)
Short sickness easy	0.02	(0.3)	0.05	(0.7)	0.03	(0.4)
Short sickness diff.	-0.23	(3.0)	-0.19	(2.6)	-0.20	(2.6)
Duration1*			-0.18	(5.2)		
Duration2		-	-0.30	(7.2)		
Duration3		-	-0.45	(6.2)		
Duration4		-	-0.62	(5.0)		
# schools	390		390		390	
# spells	8188		8188		8188	
# transitions	5272		5272		5272	
Log likelihood	-33017.71		-32965.49		-14398.10	

* Duration classes: 1:(91,182]; 2:(182,365]; 3:(365,547]; 4:(547,→)

[1] Absolute t-values in parentheses, based on the sandwich estimate of the covariance matrix of the estimator.

Specification I and II: results from concentrated likelihood with unobserved school specific effect
Specification III: results from partial likelihood with unobserved school specific effect

Table 2b. Estimation results of models with school specific fixed effects. The transition from Sickness to Work ($S \rightarrow W$)[1]

	I		II		III	
i) Variables at the individual level						
Female	-0.25	(7.3)	-0.13	(4.1)	-0.11	(3.4)
Age/ 10	0.05	(0.4)	0.10	(0.8)	0.15	**(1.1)**
(Age/ 10)^2	-0.05	(2.7)	-0.03	(1.8)	-0.04	**(2.0)**
Married	0.13	(4.0)	0.06	(2.0)	0.07	**(2.2)**
Perm . contract	-0.55	(9.0)	-0.17	(3.0)	-0.10	**(1.6)**
Part-timer	-0.09	**(2.8)**	-0.02	(0.7)	-0.03	**(1.0)**
Head	0.15	(3.7)	0.04	(1.1)	0.08	**(1.9)**
Lower groups	-0.06	(1.9)	-0.03	(1.1)	-0.05	**(1.8)**
Tenure1	-0.05	(1.4)	-0.03	(1.1)	0.004	**(0.1)**
Tenure2	-0.60	(4.4)	0.16	(1.2)	-0.09	(0.7)
Small groups ($\leq$ 20)	-0.27	**(8.2)**	-0.09	(2.9)	-0.11	(3.3)
Large groups ($\geq$ 3 1)	-0.22	**(5.0)**	-0.09	(2.3)	-0.08	(1.9)
ii) Variables at the school level						
Catholic	0.30	**(1.6)**	0.06	(0.3)	-0.05	(0.3)
# of teachers	-0.01	**(2.1)**	-0.005	(0.8)	-0.003	(0.5)
Pupil size decreasing	0.28	(5.0)	0.08	(1.6)	0.07	(1.3)
Pupil size increasing	0.12	**(2.0)**	0.05	(0.9)	0.05	(0.9)
Health services	0.01	**(0.2)**	0.06	(1.5)	0.09	(2.2)
Short replace. easy	0.27	(4.5)	0.09	(1.7)	0.09	(1.5)
Short replace. diff.	-0.28	(3.8)	-0.11	(1.5)	-0.12	(1.6)
Short sickness easy	-0.18	(2.9)	-0.03	(0.5)	-0.006	(0.1)
Short sickness diff	-0.08	**(1.2)**	-0.05	(0.8)	-0.08	(1.2)
Duration1[*]			0.04	(1.4)		
Duration2			-0.73	(18.6)		
Duration3			-1.79	(37.7)		
Duration4			-3.16	(56.3)		
# schools	419		419		419	
# spells	8092		8092		8092	
# transitions	7789		7789		7789	
Log likelihood	-28210.89		-24559.60		-19563.37	

. Duration classes: **1:(2,7]; 2:(7,14]; 3:(14,42]; 4:(42,→)**

[1] Absolute t-values in parentheses, based on the sandwich estimate of the covariance matrix of the estimator.

Specification I and II: results from concentrated likelihood with unobserved school specific effect

Specification III: results from partial likelihood with unobserved school specific effect

REFERENCES

Barmby, T.A., Orme, C.D. and J.G. Treble (1991a), Worker absenteeism: and analysis using micro data, *Economic Journal*, 101, pp.214-229

Barmby, T.A., Orme, C.D. and J.G. Treble (1991b), Analysis of worker absenteeism using discrete panel data, Working Paper Loughborough University.

Breslow, N. E. (1974), Covariance analysis of censored survival data, *Biometrics*, 30, pp. 89-99

Clayton, F.G. (1978), A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika, 65,* pp. 141-151.

Kalbfleisch, J. and R. Prentice (1980), The statistical analysis of failure time data, New *York: Wiley.*

Lindeboom, M. and M.J.M. Kerkhofs (1995), Time patterns of work and sickness absence: unobserved workplace effects in a multi-state duration model, Research memorandum Leiden University, The Netherlands.

Flinn, C.J & J.J. Heckman (1982), Models for the analysis of labour force dynamics, *Advances in Econometrics,* vol 1 pp35-95.

Gritz, M. (1993), The impact of training on the frequency and duration of employment, *journal of Econometrics,* 57,21-51.

Ham, J.C. and R. J. LaLonde (1996), The effect of sample selection and initial conditions in duration models: evidence from experimental data on training, 64, pp175-205

Lancaster, T. (1990), The econometric analysis of transition data, *Cambridge University Press.*

Ridder, G. (1984), The distribution of single spell duration data, in Neumann, G. and Westergard-Nielsen, *N (eds.), Studies in Labor Market Dynamics,* Springer Verlag, Berlin.

Ridder, G. and I. Tunali (1989), Analysis of related durations: A semi-parametric approach with an application to a study of child mortality in Malaysia, Research Memorandum Groningen University.

Ridder, G. and I. Tunali (1990), Family-specific factors in child mortality: stratified partial likelihood estimation, Research Memorandum Groningen University.

APPENDIX A.

Table Al. Means of main variables

i) Variables at the individual level	
Age	37.59
Female	0.65
Married	0.73
Tenure (years)	8.65
Permanent contract	0.89
# Hours per week	29.28
Head of the school	0.09
# pupils in class	17.99
#spells	1.67
Mean spell length	42.00
ii) Variables at the school level	
# Pupils	167.76
# Teachers	11.66
Catholic school	0.29
Protestant school	0.28
Public school	0.37
Special school	0.03
School is merged	0.25
School in big city	0.21
School in rural area	0.21
School health services available	0.55
Short term replacement easy	0.43
Short term replacement difficult	0.13
Replacement for sickness easy	0.42
Replacement for sickness difficult	0.11
# spells	19.49
Mean spell length	36.90

Table A2. Estimation results *W → Exit'*

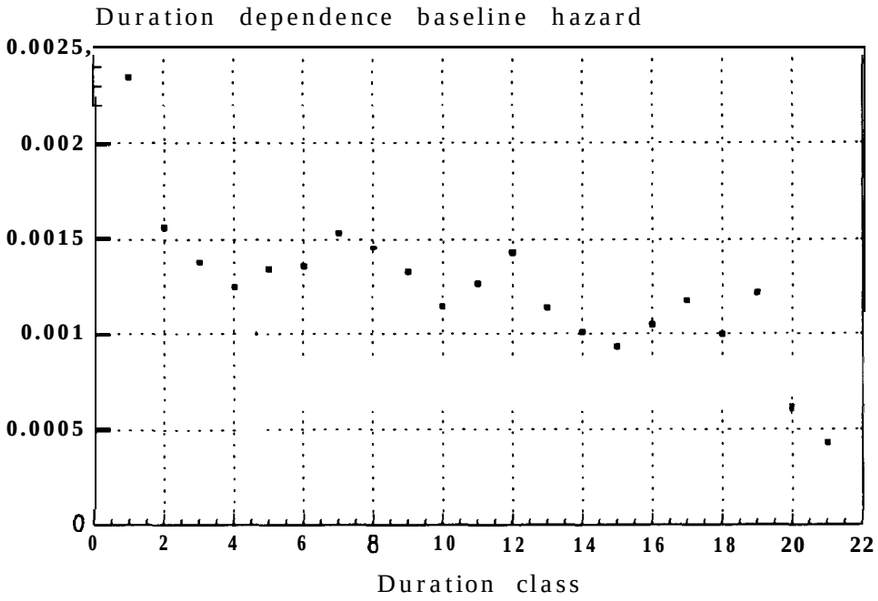
	I		II		III	
i) Variables at the individual level						
Female	-0.04	(0.3)	-0.02	(0.1)	0.08	(0.5)
Age/10	-3.02	(5.0)	-2.87	(4.7)	-2.53	(4.1)
$(Age/10)^2$	0.33	(4.3)	-0.31	(4.0)	0.28	(3.5)
Married	0.20	(1.3)	0.20	(1.3)	0.17	(1.0)
Perm . contract	-0.14	(0.7)	-0.14	(0.7)	-0.17	(0.8)
Part-timer	0.68	(4.4)	0.69	(4.5)	0.68	(4.3)
Head	0.33	(1.5)	0.34	(1.5)	0.38	(1.6)
Lower groups	-0.16	(1.1)	-0.14	(1.0)	-0.20	(1.4)
Tenure1	-0.17	(1.1)	-0.19	(1.2)	-0.23	(1.5)
Tenure2	-0.07	(0.1)	-0.22	(0.4)	-0.23	(0.4)
Small groups (≤ 20)	0.43	(2.8)	0.45	(3.0)	0.39	(2.5)
Large groups (≥ 31)	-0.20	(0.8)	-0.17	(0.7)	-0.05	(0.2)
ii) Variables at the school level						
Catholic	0.40	(0.4)	0.46	(0.5)	1.14	(1.0)
# of teachers	0.06	(1.2)	0.05	(1.1)	-0.05	(1.0)
Pupil size decreasing	-0.22	(0.8)	-0.24	(0.9)	-0.37	(1.3)
Pupil size increasing	0.08	(0.3)	0.03	(0.1)	-0.08	(0.3)
Health services	0.07	(0.4)	0.05	(0.3)	0.26	(1.4)
Short replace. easy	-0.27	(0.9)	-0.27	(0.9)	-0.55	(1.7)
Short replace. diff.	0.21	(0.7)	0.21	(0.6)	0.32	(1.0)
Short sickness easy	0.35	(1.1)	0.34	(1.1)	0.44	(1.4)
Short sickness diff.	0.09	(0.3)	0.11	(0.4)	0.12	(0.4)
Duration1*			0.47	(3.1)		
Duration2			0.96	(6.6)		
Duration3			-0.65	(2.0)		
# schools	152		152		152	
# spells	4488		4488		4488	
# transitions	308		308		308	
Log likelihood	-22546.37		-22515.33		-674.41	

* Duration classes: 1:(91,182]; 2:(182,365]; 3:(365,→)

[1] Absolute t-values in parentheses, based on the sandwich estimator of the covariance matrix of the estimator.

Specification I and II: results from concentrated likelihood with unobserved school specific fixed effect
Specification III: results from partial likelihood with unobserved school specific fixed effect

Figure Al Duration dependence in hazard work -> sick

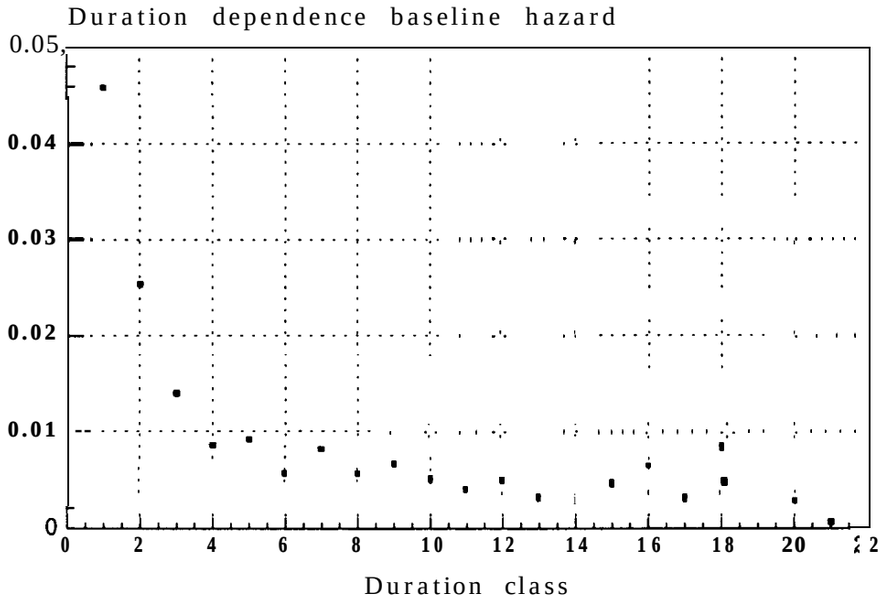


Duration classes (in days)[1]:

1	= [1,7]	12	= [127,140]
2	= [8,14]	13	= [141,170]
3	= [15,21]	14	= [171,200]
4	= [22,28]	15	= [201,230]
5	= [29,42]	16	= [231,260]
6	= [43,56]	17	= [261,290]
7	= [57,70]	18	= [291,320]
8	= [71,84]	19	= [321,365]
9	= [85,98]	20	= [366,545]
10	= [99,112]	21	= [546,730]
11	= [113,126]	22	= [731,->)

[1] Time axes are divided in a way such that the intervals contain sufficient data points to estimate the baseline hazards. Since time intervals are not equally spaced, we rescaled the baseline hazards to daily hazard rates.

Figure A2 Duration dependence in hazard sick -> work



Duration classes (in days)':

1	= [1,14]	12	= [169,189]
2	= [15,28]	13	= [190,210]
3	= [29,42]	14	= [211,231]
4	= [43,56]	15	= [232,252]
5	= [57,70]	16	= [253,280]
6	= [71,84]	17	= [281,308]
7	= [85,98]	18	= [309,336]
8	= [99,112]	19	= [337,365]
9	= [113,126]	20	= [366,456]
10	= [127,147]	21	= [457,547]
11	= [148,168]	22	= [548,->)

Time axes are divided in a way such that the intervals contain sufficient data points to estimate the baseline hazards. Since time intervals are not equally spaced, we rescaled the baseline hazards to daily hazard rates.

Table 2: Estimation results for the frictional parameters

	$1/\lambda_0$	$1/\lambda_0$	$1/\lambda_1$
All sample	162.7	15.8	126.1
Food	[159.6, 166.71 157.2	[15.2, 16.51 12.5	[116.8, 135.51 163.0
Intermediary goods	[141.1, 179.91 207.0	[10.0, 15.2] 16.5	[98.7, 245.81 162.5
Equipment	[196.5, 227.41 206.6	[13.9, 19.5] 14.6	[116.4, 194.21 174.2
Current consumption	[194.7, 223.31 153.2	[12.3, 17.11 20.8	[128.3, 214.61 199.9
Construction	[143.4, 166.61 132.5	[17.9, 23.5] 15.3	[146.7, 270.71 137.8
Trade	[126.5, 144.71 136.0	[13.5, 17.61 14.2	[101.8, 158.4] 126.2
Transport	[128.8, 147.11 215.4	[12.6, 15.9] 13.1	[95.9, 149.81 86.5
Services	[204.6, 244.41 115.7	[10.3, 16.71 13.7	[49.5, 96.7] 82.4
	[111.5, 123.61	[12.6, 15.21	[67.4, 92.9]

Time unit: month. In square brackets: the 2.5% and 97.5% percentiles of the bootstrap distribution.

Table 3: Properties of the estimated productivity distribution

	min	P_{10}	Q_1	Q_2	Q_3	P_{90}	$\frac{P_{90}}{P_{10}}$	$\frac{Q_3}{Q_1}$
All sample	**6549**	6891	7582	9021	12770	24340	3.32	1.68
Food	7056	7218	7973	9632	13440	20917	2.89	1.68
Intermediary goods	6792	7262	7920	9485	12287	19719	2.71	1.55
Equipment	7569	8092	8904	10616	14431	28487	3.52	1.62
Current consumption	7393	7565	8383	10217	16924	37089	4.90	2.01
Construction	6943	7318	8007	9386	11907	20320	2.77	1.48
Trade	6716	7090	7658	9377	13302	30436	4.29	1.73
Transport	6034	6528	7141	8296	10389	14719	2.25	1.45
Services	6267	6564	7147	8844	12424	23690	3.60	1.73

Units: French **Franc** and month. $P_{10}, Q_1, \ldots$ denote percentiles and quartiles.

Table 1: Descriptive statistics of individual data

	All sample	Food	Intermediary goods	Equipment	Current consumption	Construction	Trade	Transport telecom.	Services
Number of individuals	12214	489	1179	1361	1047	1235	1729	787	2833
Unemployed	1331	69	74	88	130	160	206	54	347
Employed	10884	420	1105	1273	917	1075	1523	733	2486
Age: mean (std deviation)	36.9(10.0)	36.0 (10.1)	38.22 (9.89)	38.0 (9.5)	37.1 (9.9)	37.4 (10.3)	35.8 (10.3)	37.9 (9.0)	35.4 (10.1)
% Women:	35.7	38.0	21.0	23.1	48.8	7.0	46.9	17.0	48.1
For unemployed:									
Transitions Unemp. → Emp.	1043	59	51	66	85	130	162	42	283
t_{ob} **censored**	190	3	6	5	14	15	22	6	53
t_{of} **censored**	288	10	23	22	45	30	44	12	64
t_{ob} **not cens. : mean (std dev)**	15.01 (16.32)	10.52 (11.87)	20.0 (18.53)	16.4 (16.7)	20.7 (19.5)	15.7 (17.6)	12.8 (14.5)	10.6 (12.5)	13.5 (15.2)
$t_{of\ not}$ **cens. : mean (std dev)**	4.10 (5.44)	3.03 (4.53)	3.81 (5.40)	3.9 (5.2)	4.9 (5.7)	4.1 (5.2)	3.9 (5.3)	4.6 (6.6)	3.9 (5.5)
# observed accepted wages	190	7	6	15	20	46	26		42
For employed:									
Transitions Emp. → Emp.	528	19	38	38	37	80	74	33	164
Transitions Emp. → Unemp.	812	26	73	81	77	111	110	26	251
t_{1b} **Censored**	155	2	14	10	15	20	23	9	49
t_{1f} **censored**	9544	375	994	1154	803	884	1339	674	2071
$t_{1b\ not}$ **cens. : mean (std dev)**	111.8 (103.9)	117.82 (103.13)	139.0 (114.8)	143.3 (111.2)	110.3 (100.8)	99.5 (96.2)	94.2 (95.6)	131.8 (108.0)	78.2 (88.0)
$t_{1f\ not}$ **cens. : mean (std dev)**	10.35 (7.05)	9.44 (6.59)	11.1 (7.1)	10.9 (7.2)	10.0 (7.1)	11.3 (7.1)	10.5 (7)	10.1 (7.0)	9.9 (7.0)
# observed wages	10161	396	1075	1237	869	1026	1408	711	2226
Cross-sectional wages':									
minimum	4497	4500	4500	4500	4500	4500	4500	4500	4497
P_{10}	5000	4836	5158	5405	4700	5000	4918	5612	5000
Q_1	5850	5500	6000	6300	5250	5700	5580	6500	5694
Q_2	7200	6700	7256	7800	6500	6631	6808	7750	7042
$Q3$	9694	8667	9185	10500	9208	8125	9225	9750	9898
P_{90}	13650	10933	12000	15000	13612	10761	13700	13000	14000
P_{90}/P_{10}	2.73	2.26	2.32	2.77	2.89	2.15	2.78	2.31	2.80
Q_3/Q_1	1.65	1.57	1.53	1.66	1.75	1.42	1.65	1.50	1.73
mean (std deviation)	8468 (3992)	7837 (3885)	8313 (3727)	9135 (4213)	8152 (4302)	7538 (3115)	8195 (3946)	8743 (3645)	8440 (4070)

Units: French Franc and month.

1: $Q_1, Q_2, Q_3, P_{10}, P_{90}$ **are respectively the first, second and third quartile, and the tenth and ninetieth percentile of the cross-sectional wage distribution.**

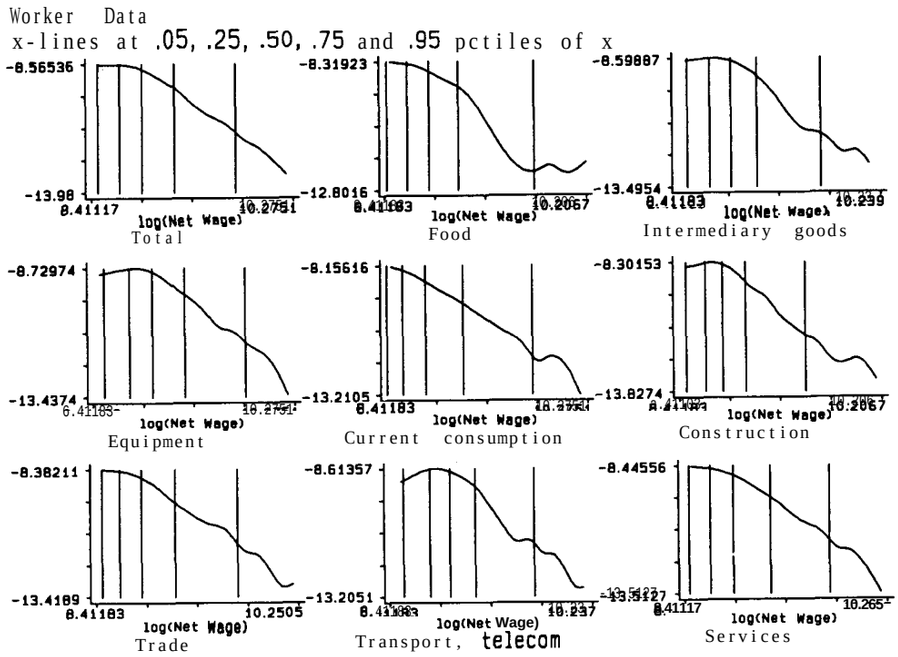


Fig. 5: Earnings Log-Density

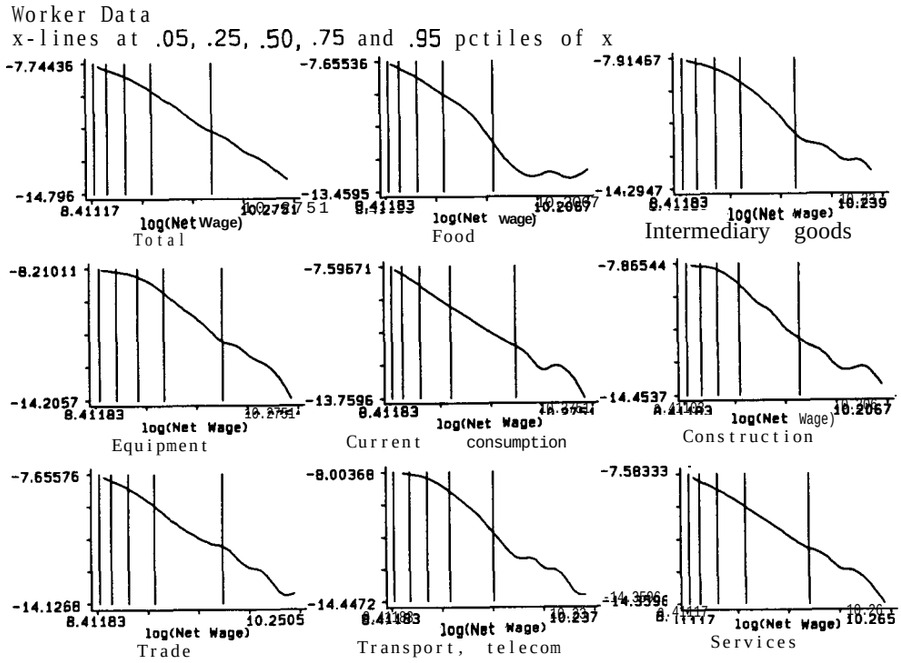


Fig. 6: Wage Offer Log-Density

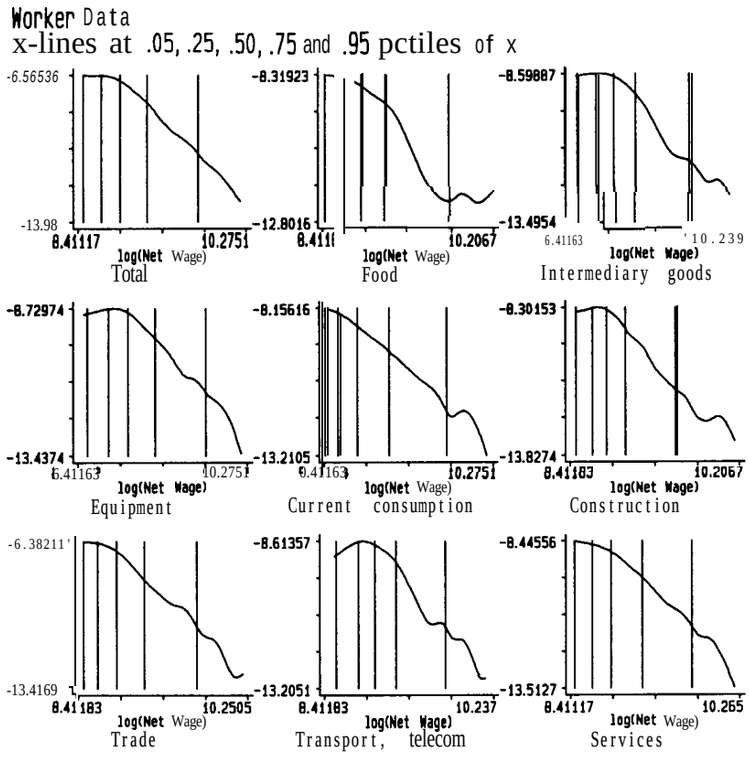


Fig. 5: Earnings Log-Density

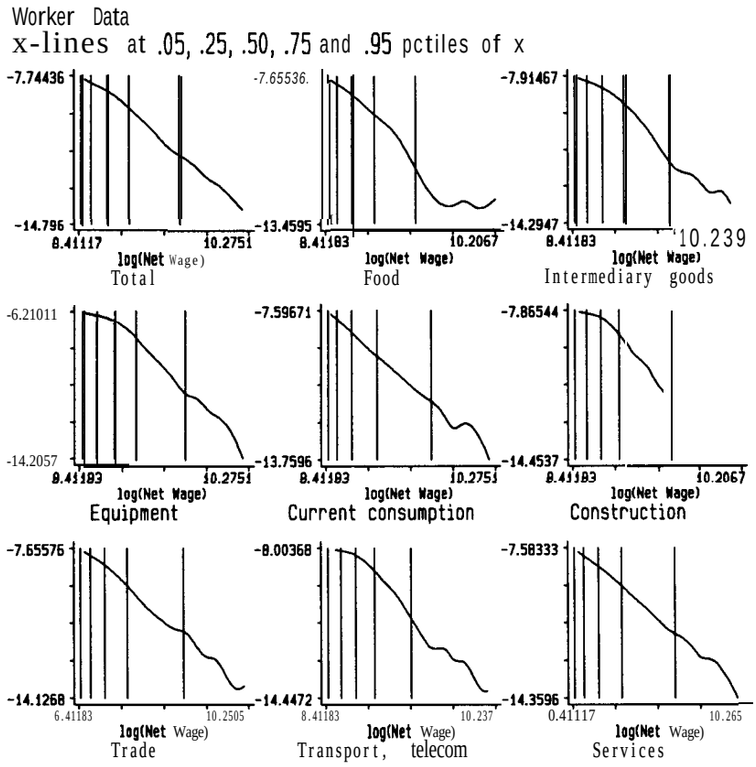


Fig. 6: Wage Offer Log-Density

Table 4: 'Descriptive statistics of firm data

	Food	Intermediary goods	Equipment	Current consumption	Construction	Trade	Transport telecom.	Services
Number of enterprises	3784	8928	7324	9562	10586	20438	5446	22514
Wage cost per worker:								
P_{10}	**9527**	11352	11480	8848	11293	**9345**	10166	9126
Q_1	11381	12995	13427	10844	12583	11513	11947	11812
Q_2	13490	15028	15743	13576	14524	14611	14034	15453
Q3	16149	17526	18876	17513	17076	18701	16637	21020
P_{90}	19450	20593	22333	22948	19857	24642	20615	29467
P_{90}/P_{10}	1.63	1.81	1.94	2.55	1.75	2.63	2.02	3.22
Q_3/Q_1	1.41	1.34	1.40	1.61	1.35	1.62	1.39	1.77
Value-added per worker:								
P_{10}	10738	12636	12456	9189	12128	10948	11157	9699
Q_1	13798	15769	15509	12619	14279	14636	14497	13844
Q_2	18210	19627	19353	17307	17007	19620	17977	19286
Q3	25638	24964	24279	23754	2042 1	26760	22493	26848
P_{90}	36557	32456	3078 1	33283	24550	39308	28595	39662
P_{90}/P_{10}	3.40	2.56	2.47	3.62	2.02	3.59	2.56	4.08
Q_3/Q_1	1.85	1.58	1.56	1.88	1.43	1.82	1.55	1.93
Monopsony power:								
P_{10}	**0.00**	**0.00**	-0.07	-0.05	-0.01	-0.04	-0.02	-0.07
Q_1	0.14	0.12	0.08	0.08	0.07	0.13	0.10	0.07
Q_2	0.26	0.22	0.17	0.19	0.13	0.25	0.21	0.17
Q_3	0.40	0.34	0.28	0.30	0.20	0.37	0.31	0.29
P_{90}	0.54	0.46	0.38	0.42	0.28	0.49	0.41	0.44
Employment:								
P_{10}	20	21	21	21	20	7	20	12
Q_1	25	26	26	26	23	20	25	22
Q_2	39	39	41	39	32	28	36	32
Q3	81	75	89	74	49	45	60	54
P_{90}	195	185	233	171	98	91	136	119

Units: French Franc and month. $P_{10}, Q_1, \ldots$ **denote percentiles and quartiles** the variable un er consideration.

Table' 4: Descriptive statistics of firm data

	Food	Intermediary goods	Equipment	Current consumption	Construction	Trade	Transport telecom.	Services
Number of enterprises	3784	8928	7324	9562	10586	20438	5446	22514
Wage cost per worker:								
P_{10}	9527	11352	11480	8848	11293	9345	10166	9126
Q_1	11381	12995	13427	10844	12583	11513	11947	11812
Q_2	13490	15028	15743	13576	14524	14611	14034	15453
Q_3	16149	17526	18876	17513	17076	18701	16637	21020
P_{90}	19450	20593	22333	22948	19857	24642	20615	29467
P_{90}/P_{10}	1.63	1.81	1.94	2.55	1.75	2.63	2.02	3.22
Q_3/Q_1	1.41	1.34	1.40	1.61	1.35	1.62	1.39	1.77
Value-added per worker:								
P_{10}	10738	12636	12456	9189	12128	10948	11157	9699
Q_1	13798	15769	15509	12619	14279	14636	14497	13844
Q_2	18210	19627	19353	17307	17007	19620	17977	19286
Q3	25638	24964	24279	23754	2042 1	26760	22493	26848
P_{90}	36557	32456	30781	33283	24550	39308	28595	39662
P_{90}/P_{10}	3.40	2.56	2.47	3.62	2.02	3.59	2.56	4.08
Q_3/Q_1	1.85	1.58	1.56	1.88	1.43	1.82	1.55	1.93
Monopsony power:								
P_{10}	0.00	0.00	-0.07	-0.05	-0.01	-0.04	-0.02	-0.07
Q_1	0.14	0.12	0.08	0.08	0.07	0.13	0.10	0.07
Q_2	0.26	0.22	0.17	0.19	0.13	0.25	0.21	0.17
Q_3	0.40	0.34	0.28	0.30	0.20	0.37	0.31	0.29
P_{90}	0.54	0.46	0.38	0.42	0.28	0.49	0.41	0.44
Employment:								
s o	20	21	21	21	20	7	20	12
Q_1	25	26	26	26	23	20	25	22
Q_2	39	39	41	39	32	28	36	32
Q3	81	75	89	74	49	45	60	54
P_{90}	195	185	233	171	98	91	136	119

Units: French Franc and month. $P_{10}, Q_1, \ldots$ denote percentiles and quartiles the variable under consideration.